Predição de Função de Proteínas Através da Extração de Características Físico-Químicas

Thiago Assis de Oliveira Rodrigues [1]
Larissa Fernandes Leijôto [1]
Poliane C. Oliveira Brandão [1]
Cristiane Neri Nobre [1]

Data submissão: 29.07.2014
Data aceitação: 16.02.2015

Resumo: Com a conclusão do projeto Genoma, o número de novas proteínas descobertas tem crescido, mas devido ao alto custo e da demora dos processos de descoberta da função de proteínas, apenas uma pequena parcela das mesmas tem sua função conhecida. Este trabalho apresenta uma metodologia para predição de função de proteínas através da extração de características de suas estruturas, presentes no banco de dados Sting_DB, a utilização da Transformada Discreta do Cosseno, a codificação da estrutura primária, o balanceamento de classes e a utilização de Máquinas de Vetores de Suporte. Os valores médios obtidos para a precisão, sensibilidade, acurácia e especificidade foram, respectivamente de 80%, 71%, 74% e 77%. Os resultados foram comparados com outros trabalhos da literatura, e mostraram um aumento de 10% na taxa de precisão.

Abstract: With the conclusion of the Genome project, the number of new discovered proteins has grown, but due to the high cost and delay of the processes of protein function discovery, just a small parcel of them have their function known. This article presents a methodology for the prediction of protein function through the extraction of the characteristics of its structure, which are present in Sting_DB database, the Discrete Cosine Transform, the encoding of primary structure, class balancing and the use of Support Vector Machines. The average values obtained for precision, sensitivity, accuracy and specificity were 80%, 70%, 74% and 77%, respectively. The results were compared with other studies in the literature and showed an increase of 10% in the accuracy rate.

[1] Instituto de Ciências Exatas e Informática
Pontifícia Universidade Católica de Minas Gerais
Belo Horizonte, Brasil
{taorodrigues,larissa.leijoto}@sga.pucminas.br
{polianecoliveira@gmail.com,nobre@pucminas.br}

1 Introdução

A bioinformática tem por objetivo o estudo e aplicação de técnicas computacionais a diversas áreas da biologia [1]. Nesse contexto, a computação pode ser aplicada na resolução de uma série de problemas, tais como: comparação de sequências (DNA, RNA e proteínas), montagem de fragmentos, reconhecimento de genes, identificação e análise da expressão de genes, reconstrução de árvores filogenéticas e determinação da estrutura e função de proteínas. Dentre esses problemas, a predição da função de proteínas é um grande desafio da bioinformática. O estudo relativo a esse assunto vem caracterizando uma nova fase para as pesquisas genéticas, denominada proteômica [2], que envolve a identificação de todas as proteínas expressas pelo genoma, bem como a determinação de suas funções fisiológicas e patológicas.

As proteínas são as macromoléculas utilizadas como matéria-prima e são componentes funcionais das células, sendo a segunda maior fração em peso, perdendo apenas para a água. Elas são muito diversificadas e, por isso, apresentam várias formas de classificação. Em geral, podemos classificá-las de acordo com suas quatro etapas estruturais: primária, secundária, terciária e quaternária.

Com a finalização do Projeto Genoma, o número de proteínas depositadas em bancos de dados tem crescido. Em [3] os autores fizeram um levantamento da quantidade de proteínas cuja função foi descoberta pelos principais laboratórios do mundo ao longo dos anos de 2.000 e 2.010 (Fig. 1).

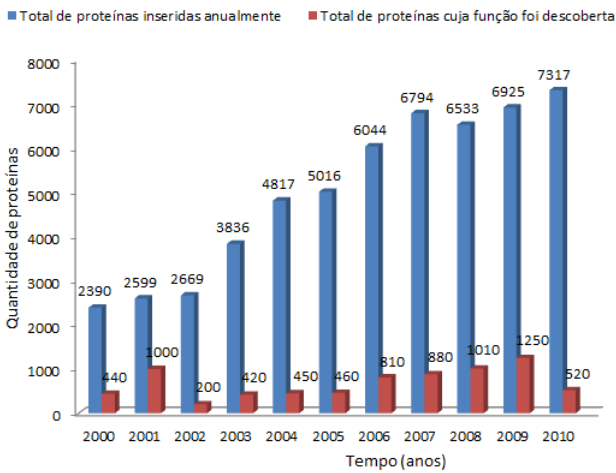


Figura 1. Comparação entre o crescimento anual das proteínas no PDB e quantidade de proteínas com função conhecida entre os anos 2000 e 2010.

Nota-se que a quantidade de proteínas adicionadas anualmente ao *Protein Data Bank* (PDB) [4] varia de 7 a 8 mil, enquanto que as proteínas com função conhecida, não chegam a 1.500 por ano. Atualmente o PDB já possui mais de 80.000 proteínas. É importante destacar que muitas dessas proteínas depositadas no PDB podem ser homólogas entre si. Com isso, a identificação da função de uma proteína poderia possibilitar a definição da função de muitas outras.

Esse enorme crescimento do número de proteínas acarreta em um grande desafio do ponto de vista laboratorial e computacional. Existem métodos tradicionais para determinar a estrutura, como por exemplo: Cristalografia por Difração de Raios-X e a Ressonância Nuclear Magnética [5]. Contudo, nem sempre é possível determinar a função de uma proteína somente pela sua estrutura. As enzimas que são proteínas catalisadoras, ou seja, aceleram reações bioquímicas possuem afinidades e especificidades a substratos diferentes, exercendo diversas funções em uma célula. Por isso, não só a estrutura, mas também sua sequência, e principalmente os aminoácidos que formam o seu sítio ativo, apresentam um papel importante na definição de sua função. Do ponto de vista computacional, um dos maiores desafios é selecionar as características realmente relevantes para que ocorra uma previsão da função de proteínas com exatidão, reduzindo assim a necessidade de testes laboratoriais.

Este trabalho tem o objetivo de propor uma metodologia de extração de características de proteínas para predição de sua função, dada de acordo com a classe a que pertencem, a partir do banco de dados público PDB [4]. A predição da função será baseada aos aspectos estruturais da proteína extraídos do banco de dados Sting_DB [**?**].

É importante ressaltar que nos limitaremos a utilizar um subgrupo das proteínas, citado por [6], que utilizou as seguintes superfamílias: Oxidoredutases, Transferases, Hidrolases, Liases, Isomerases e Ligases.

Este artigo está dividido da seguinte maneira: na Seção 2 são detalhados os principais conceitos utilizados neste trabalho, e que são primordiais para o entendimento do mesmo, na Seção 3 são apresentados os trabalhos relacionados, a Seção 4 explicita as questões de pesquisa deste trabalho, a Seção 5 detalha a metodologia proposta, na Seção 6 os resultados dos experimetos são exibidos e a Seção 7 traz as considerações finais deste trabalho.

2 Conceitos Básicos

2.1 Proteínas

As proteínas são os compostos orgânicos mais comuns em um organismo, os mais abundantes depois da água e também os de maior variedade molecular [7]. Estão presentes em todas as estruturas celulares, desde a membrana até o núcleo, compondo as substâncias intercelulares, hormônios, anticorpos, dentre outros.

As reações químicas que ocorrem no interior dos aminoácidos formam uma proteína e determina como essa proteína está organizada tridimensionalmente. Para compreender as propriedades de uma proteína, é necessário descrever como os aminoácidos são formados.

Sua estrutura é constituída por um carbono central ou carbono alfa, que se liga a quatro grupos: o grupo amina (NH2), o grupo carboxílico (COOH), um hidrogênio (H) e uma cadeia que é denominada cadeia lateral, responsável por diferenciá-los [8]. Proteínas possuem quatro níveis de organização: primária, secundária, terciária e quaternária.

- **Estrutura Primária** é a sequência de aminoácidos ao longo de sua cadeia.
- **Estrutura Secundária** consiste na relação espacial entre os aminoácidos que estão próximos na estrutura primária. Nas proteínas, as unidades básicas da estrutura secundária são as alfa hélices e as folhas beta.
- **Estrutura Terciária** é como os átomos de uma cadeia polipeptídica estão organizados em espaço tridimensional.
- **Estrutura Quaternária** são as interações entres as diversas cadeias de aminoácidos presentes em uma proteína.

As enzimas são proteínas catalisadoras, que aceleram a velocidade das reações bioquímicas. A *International Union of Biochemistry and Molecular Biology* (IUBMB) classificou as enzimas em seis grandes classes de proteínas, de acordo com o tipo de reação que catalisam, e atribuiu a cada classe, um número de classificação, conhecido por *Enzyme Commision* (EC). A Tabela 1 apresenta o número de EC, o nome e as principais funções de cada uma dessas classes. Este artigo faz a predição da função destas enzimas.

Tabela 1. Classes de enzimas e suas respectivas funções.

EC	Classe	Função
1	Oxidoredutases	Catalisam reações de oxidação ou de redução.
2	Transferases	Transferência de grupos funcionais entre duas moléculas.
3	Hidrolases	Reações de hidrólise de várias ligações covalentes.
4	Liases	Quebra de ligações covalentes e remoção de moléculas de água, amônia e gás carbônico.
5	Isomerases	Modificações de uma única molécula, sem partição de outra.
6	Ligases	Reações de formação de uma nova molécula a partir da ligação entre outras duas.

Fonte: Adaptado de [4].

2.2 Support Vector Machines

As *Support Vector Machines* (SVMs) foram desenvolvidas por Vapnik e Vladimir [9] e constituem uma técnica de aprendizado baseada no fato de que, em altas dimensões do espaço de características, todos os problemas se tornam linearmente separáveis. Algumas das principais características que tornam seu uso atrativo, segundo [10], são: boa capacidade de generalização, robustez em grandes dimensões e teoria bem definida.

As SVMs utilizam funções denominadas de *kernels*, capazes de mapear um conjunto de dados em diferentes espaços, possibilitando a utilização dos hiperplanos. Quatro tipos de *kernel* são frequentemente utilizados em SVMs: Linear, Polinomial, Gaussiano (*Radial-Basis Function*) e Sigmoidal. Cada *kernel* [10] possui a sua equação e seus respectivos parâmetros, apresentados na Tabela 2, e é altamente sensível a pequenas variações desses.

Tabela 2. *Kernels* comumente utilizados na SVM.

Tipo da Função	Equação	Parâmetros
Linear	$X_i^T X_j$	-
Polinomial	$(\gamma X_i^T X_j + r)^d$	r, d
Gaussiano(RBF)	$exp(\gamma \|X_i - X_j\|^2)$	γ
Sigmoidal	$\tanh(\gamma X_i^T X_j + r)$	γ, r

Fonte: Hsu et al. [11]

2.3 Bases de dados

O conjunto de proteínas utilizado neste trabalho foi o mesmo utilizado por [6], [12] e [13]. As proteínas foram extraídas do PDB, o maior e mais completo repositório de proteínas existente.

Outro banco de dados utilizado foi o Sting_DB [**?**], desenvolvido pelo laboratório de Biologia Computacional da Embrapa-Brasil. Esse banco de dados processa automaticamente proteínas no formato PDB e disponibiliza diversas características das mesmas.

Entre os módulos do Sting_DB, há o *Java Protein Dossier* (JPD) [14], uma ferramenta que fornece aos usuários uma vasta coleção de parâmetros físico-químicos descrevendo a estrutura da proteína, estabilidade e interações com outras macromoléculas.

Ao mesmo tempo, o JPD é um passo em direção à criação de uma base de dados diversificada de descritores de estrutura e função, que podem ser usados como uma plataforma para a aquisição de novos conhecimentos. Além disso, ele permite que sejam salvos os dados de todos os seus parâmetros para cada aminoácido presente em uma determinada cadeia de proteína.

Todas as características utilizadas neste trabalho foram extraídas deste repositório, e podem ser agrupadas em:

- **Evolutivas**, calculadas por meio da mudança das proteínas, ou seja, o quanto as sequências evoluíram ao longo do tempo.
- **Contatos inter-atômicos**, obtidos por meio do contato entre os átomos presentes em cada resíduo da proteína.
- **Físico-químicas**, obtidas por meio das atrações ocorridas pelos diversos tipos de ligações entre os aminoácidos.
- **Estrutura geométrica**, calculadas por meio da estrutura tridimensional da proteína.
- **Superfície**, calculada através das cavidades contidas na superfície de uma proteína, de onde os ligantes se acoplam.

3 Trabalhos Relacionados

Os autores de [6] apresentaram um método para predição de classes de proteínas a partir de seus dados estruturais, utilizando atributos simples, tais como o conteúdo da estrutura secundária, propensões de aminoácidos e propriedades de superfície. A ideia do método proposto é, a partir de um grupo de atributos estruturais, classificar cada proteína dentro de uma das seis classes protéicas. Eles utilizaram o banco de dados ASTRAL SCOP [15] e obtiveram uma precisão de 35% usando SVMs.

No trabalho de [12], o objetivo foi melhorar o processo de seleção de parâmetros para aumentar a precisão do modelo de classificação de proteínas. Por meio de uma abordagem híbrida que utilizou recursos da matemática e da estatística, os autores utilizaram um conjunto de proteínas e abordaram três desafios presentes na classificação de proteínas: o ruído presente nos parâmetros, o grande número de variáveis e o número não balanceado de membros por classe.

Os autores utilizaram o banco de dados Sting_DB e utilizaram a transformada discreta do cosseno (TDC), para eliminar o problema de se trabalhar com tamanhos diferentes de cadeias de aminoácidos. Depois, efetuaram um balanceamento entre as classes de proteínas utilizadas, pois o fato de uma classe possuir muito mais elementos do que a outra pode afetar a precisão do classificador. Para diminuir o número de variáveis utilizadas, os autores aplicaram a técnica de *Principal Component Analysis* (PCA). E por último, utilizaram quatro algoritmos de aprendizado de máquina a partir da ferramenta WEKA [16]: Árvore de Decisão, Modelo Bayesiano, Redes Neurais e SVM.

Os primeiros testes foram realizados utilizando ao todo 241 variáveis, sem a aplicação da técnica de PCA, e obteve-se uma taxa de 68,41% de precisão para o método de Árvore de Decisão, seguido do método de Redes Neurais com 64,49%, SVM com 42,60% e método Bayesiano com 41,42%. Posteriormente, foram realizados testes utilizando a redução de variáveis, que reduziu o número de parâmetros para 80. Nesse caso, o método baseado em Árvore de Decisão obteve novamente o melhor resultado com precisão igual a 69,01%,

seguido do método de Redes Neurais com 60,35%, SVM com 59,17% e o método Bayesiano com 49,11% de precisão.

Em [17], foi discutida a questão do pré-processamento dos dados antes da etapa de classificação. Foram analisadas duas metodologias de codificação do alfabeto que representa os aminoácidos com o intuito de verificar qual delas apresentaria o menor tempo de processamento. A primeira metodologia é chamada de $n-gram$, e é utilizada em indexação de textos para efetuar o casamento de padrões. Ela verifica quantas vezes uma determinada sequência de caracteres $(1..n)$ ocorre em todo o texto, além de verificar a frequência dos aminoácidos em cada proteína. A segunda metodologia utilizou a escala de hidrofobicidade para criar uma codificação para os aminoácidos dividindo-os em três categorias: hidrofóbicos, neutros e hidrofílicos. Com base no tempo de execução de cada metodologia observou-se que a escala de hidrofobicidade foi a mais eficiente.

[13] utilizaram as mesmas proteínas analisadas em [6] e fez a predição da sua função a partir de características primárias e secundárias. No entanto, uma vez que as proteínas possuem tamanhos diferentes, foi necessário criar uma metodologia que mantivesse as proteínas com o mesmo tamanho. Isto é necessário, pois os autores utilizaram o classificador SVM e este precisa receber entradas de tamanhos iguais para a sua classificação. A solução encontrada foi adicionar zeros na sequência até que as cadeias atingissem o tamanho da maior cadeia presente no banco de dados. Depois, foi feita uma normalização dos dados, com base nos menores e maiores valores utilizados, juntamente com uma codificação que utilizou tabelas *hash*. Por fim, foi utilizado um algoritmo genético para detectar os melhores valores para os parâmetros do classificador SVM. Com a metodologia proposta, os autores obtiveram acurácia, sensibilidade, precisão e especificidade de 79,74%, 70,31%, 70,06% e 96,72%, respectivamente.

[18] propõe um modelo de predição de função de proteínas, por meio de SVM, utilizando suas funções moleculares, como parâmetros estruturais, calculados a partir da conformação espacial da própria proteína retirada do Sting_DB. O modelo utilizou as mesmas características propostas por [12]. Foi criado um classificador para cada função escolhida, e caso a proteína executasse a função desejada, ele devolveria uma resposta do tipo sim ou não. Em seguida, foi criado um classificador global, que reúne todas as funções trabalhadas. O autor utilizou as métricas de precisão e sensibilidade que obtiveram, respectivamente, 98% e 93%.

A Tabela 3 apresenta um sumário das técnicas utilizadas em cada um dos trabalhos relacionados.

Tabela 3. Síntese das metodologias dos trabalhos relacionados.

Metodologia	**Trabalhos relacionados**				
	Dobson e Doig [6]	Oliveira [12]	Rossi e Brunetto [17]	Dias [18]	Resende [13]
Banco de dados	Astral Scop	PDB Sting_DB	PDB	Sting_DB	PDB
Classificador	SVM	Árvore de decisão Modelo Bayesiano Rede Neural SVM	-	SVM	SVM
Codificação	-	-	n-grams Hidrofobicidade	-	Tabela *hash*
Padronização do tamanho vetor	-	TDC PCA	-	-	Inserção de zeros
Frequência de aminoácidos	-	✗	✗	-	-
Balanceamento	-	✗	-	-	-
Normalização	-	-	-	-	✗

- Significa que a técnica não foi utilizada ou que os autores não afirmaram sobre o seu uso.
✗ Significa que a técnica foi utilizada.

4 Questões de pesquisa

Determinar a sequência de uma proteína é relativamente mais fácil do que determinar a sua função, o que leva a uma grande diferença entre o número de sequências e o número de proteínas com suas funções conhecidas. Logo deseja-se saber, qual a forma mais viável de se extrair características de proteínas conhecidas, utilizando métodos computacionais, para classificarmos proteínas com função desconhecida. Desta forma, nós focamos nas seguintes questões de pesquisa:

1. Que metodologia utilizar para predizer a função de uma proteína a partir de suas características físico-químicas?
2. Que outras características existentes podem melhorar o desempenho do classificador?
3. Qual o impacto do balanceamento na predição da função das proteínas consideradas?

5 Metodologia

Esta seção descreve a metodologia proposta. O diagrama da Figura 2 ilustra todas as etapas presentes nesta metodologia, dividida em três grandes partes: pré-processamento dos dados (verde), processamento (laranja) e avaliação (azul).

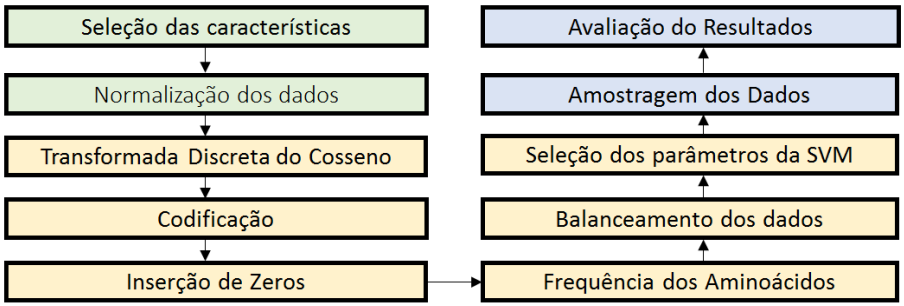


Figura 2. Fluxograma da metodologia proposta.

5.1 Seleção das principais características

Algumas características utilizadas neste trabalho são baseados em [12] e [18]. No trabalho de [12] foram utilizadas 11 características das 18 existentes no banco de dados Sting_DB. Já o trabalho de [21] utilizou apenas 9 dessas 11 características, pois através de alguns testes verificou-se que havia redundância das informações entre 2 delas. Desde a data de publicação destes trabalhos, novas características foram inseridas no banco de dados, sendo que atualmente existem 27 características diferentes distribuídas entre os grupos descritos na Seção 2.6.

Porém, a maioria dessas características possui diversos parâmetros que, ao serem alterados, geram novos valores que são totalmente diferentes uns dos outros. Ao todo, possuímos 338 características geradas através da variação dos parâmetros das 27 características principais.

Foram realizados diversos testes e, dentre todas as 338 caraterísticas, as 10 selecionadas para serem utilizadas neste trabalho estão listadas abaixo, sendo que as 6 primeiras são as mesmas utilizadas por [12] e [18] e as quatro últimas foram escolhidas pelos autores deste trabalho.

- **Potencial Eletrostático:** Os átomos que compõem os resíduos de aminoácido das proteínas podem, em determinadas condições, apresentar carga elétrica, que interagem com outras regiões carregadas da própria proteína ou ainda com outras moléculas e/ou íons de seu ambiente. Portanto, em um determinado ponto do espaço é possível calcular o potencial eletrostático devido a cargas presentes nas macromoléculas ao redor do ponto. Neste trabalho, foi utilizada a média do potencial eletroestático.

- **Hidrofobicidade:** Proteínas que possuem em sua cadeia lateral átomos de nitrogênio e oxigênio capazes de estabelecerem ligações de hidrogênio com as moléculas de água são chamadas de hidrofílicas. Os aminoácidos hidrofílicos são divididos em três cate-

gorias: Polares sem carga (neutros), polares com carga negativa (ácidos) e polares com carga positiva (básicos). Ao restante é associado o termo hidrofóbico.

- **Ordem de *Cross Link***: Devido ao dobramento da sequência de resíduos de aminoácido na estrutura tridimensional, resíduos são colocados próximos no espaço, e podem, portanto, interagir entre si. O parâmetro *Cross Link* presente no JPD leva essa característica em conta, e é definida em relação ao número dos contatos estabelecidos entre seguimentos de resíduos de aminoácido de, no mínimo, 15 resíduos na estrutura primária da proteína.

- **Ordem de *Cross Presence***: Seguindo a mesma definição de Cross Link, esse valor é calculado através da contagem de todos os resíduos de aminoácidos que estão dentro da sonda esférica, mesmo que os resíduos de aminoácido não estejam estabelecendo nenhum contato entre si.

- **Densidade:** A densidade local de cada resíduo de aminoácido é calculada utilizando uma abordagem de sonda esférica, onde as massas dos átomos internos à sonda esférica são somadas e divididas pelo volume da sonda esférica.

- **Distância do Centro de Gravidade:** Representa a distância entre o carbono alfa (CA) de cada resíduo e o centro de massa da cadeia (baricentro).

- **Esponjicidade:** A esponjicidade é uma medida do espaço vazio do nano-ambiente de cada aminoácido e segue a mesma abordagem descrita para a densidade.

- **Ocupação Múltipla:** A presença de dois ou mais conjuntos de coordenadas para o mesmo átomo no arquivo PDB é devido à interpretação do mapa de densidade de elétrons onde o experimento registra uma difração a partir dos cristais de congelamento da mesma molécula, mas com as diferentes posições de espaço para um determinado aminoácido.

- ***Hot−Spots*:** Esta característica indica a existência de cavidades hidrofóbicas nas superfícies das proteínas. Elas são potencialmente importantes para a identificação de porções superficiais que podem se envolver nas interações das proteínas.

- **Curvatura:** Uma superfície de uma proteína não é uma superfície plana. Nela estão presentes áreas côncavas e convexas. As porções côncavas das superfícies podem ser indicadores de prováveis interações com outras proteínas. Durante o processo de ancoragem, duas moléculas devem encaixar geometricamente porções opostas das superfícies. A curvatura média dessa porção de superfície deve ser igual a duas superfícies, de sinais opostos.

5.2 Normalização

O propósito da normalização é minimizar os problemas oriundos do uso de unidades e dispersões distintas entre as variáveis. Isso evita que uma dimensão se sobreponha em relação às outras, prevenindo assim que o aprendizado fique estagnado. A técnica de normalização utilizada neste trabalho foi a $MaxMin\ Equalizada$ (Equação 1), que utiliza os valores máximo e mínimo para normalizar linearmente os dados entre $[0, 1]$.

$$novo\_x = \frac{x - min(x)}{max(x) - min(x)} \tag{1}$$

onde $novo\_x$ é o novo valor de x para um determinado número x que será normalizado, $min(x)$ é o menor valor de x presente nos dados e $max(x)$ é o maior valor de x.

Para que a aplicação da normalização fosse possível, foi feito um pré-processamento dos dados, de forma a obter os maiores e menores valores para cada uma das características extraídas das proteínas. Posteriormente, foi feita a normalização dos dados por característica.

5.3 Transformada Discreta do Cosseno

Para que a utilização de um classificador seja possível, o tamanho de todos os vetores de entrada deve ser o mesmo. No entanto, a diferença entre a quantidade de aminoácidos de cada proteína faz com que cada vetor tenha um comprimento diferente. A Figura 3 ilustra a variação do tamanho das proteínas no contexto deste trabalho.

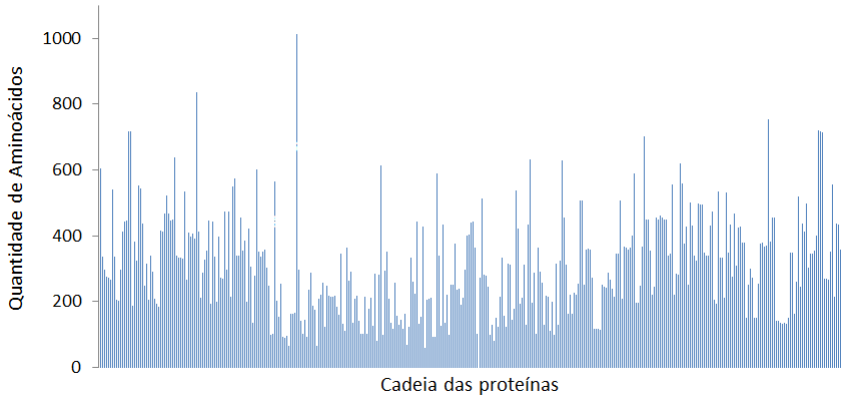


Figura 3. Quantidade de aminoácidos entre as cadeias das proteínas consideradas.

Para solucionar este problema foi utilizada a técnica da Transformada Discreta do Cosseno (TDC) [19], também utilizada nos trabalhos de [12] e [18]. A TDC foi escolhida pois é uma transformação que preserva as normas e os ângulos dos vetores e é uma transformada

para números reais, ao contrário da Transformada Discreta de Fourier, definida sobre o corpo dos números complexos. Esta técnica representa um vetor com os dados originais, Figura 4 (a), em um sinal no domínio da frequência, Figura 4 (b). Feito isso, a TDC faz com que os valores mais significativos fiquem localizados nos primeiros coeficientes do vetor transformado, ao passo que nos últimos coeficientes ficam os ruídos.

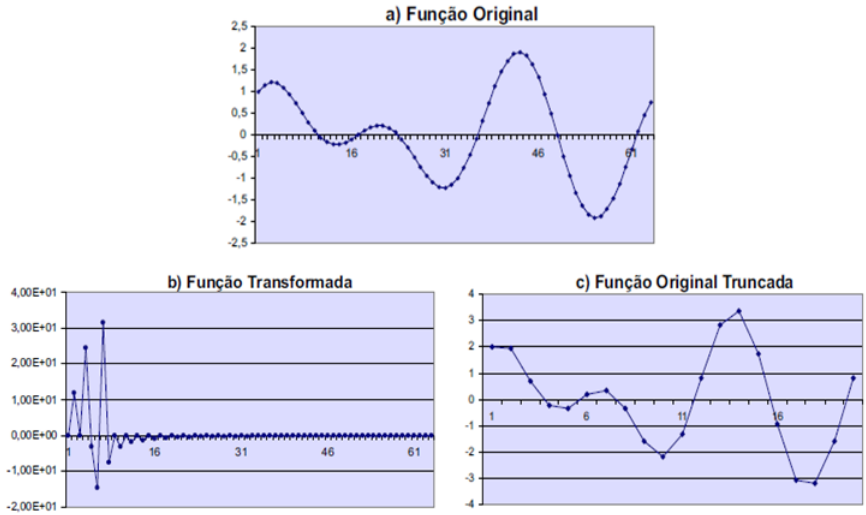


Figura 4. Em (a) temos a função original. Após aplicarmos a TDC, podemos visualizar em (b) a função no domínio das frequências. Aplicando-se a TDIC percebemos que o comportamente da curva em (c) é o mesmo da função original.

Ao selecionarmos apenas os primeiros n coeficientes, estamos dizendo que a parte mais relevante da informação foi obtida. Para verificar quais são os valores originais correspondentes deve-se usar a Transformada Inversa do Cosseno (TDIC). O resultado serão valores próximos aos valores originais, pois o comportamento original da curva é preservado, como pode ser visto na Figura 4 (c).

5.4 Codificação

As estruturas das proteínas são fonte de muita informação que pode ser utilizada durante o processo de classificação. No caso da estrutura primária das proteínas, que consiste na disposição dos aminoácidos em suas cadeias, deve ser utilizada alguma técnica de codificação que permita o mapeamento das informações das estruturas para a forma numérica. No trabalho de [20] foi proposta uma codificação da estrutura primária das proteínas, que utiliza os valores da escala de Hidrofobicidade proposta por Kyte e Doolittle [21]. Essa codificação,

apresentada na Tabela 4, foi utilizada nesse trabalho e acrescentada ao vetor de características de cada proteína.

Tabela 4. Valores de Hidrofobicidade e codificação adotada em valores reais.

Aminoácido (símbolo)	Valor K.D	Valor Real Categoria	Classificação
I	+4,5	0,05	Hidrofóbico
V	+4,2	0,10	Hidrofóbico
L	+3,8	0,15	Hidrofóbico
F	+2,8	0,20	Hidrofóbico
C	+2,5	0,25	Hidrofóbico
M	+1,9	0,30	Hidrofóbico
A	+1,8	0,35	Hidrofóbico
G	-0,4	0,40	Neutro
T	-0,7	0,45	Neutro
S	-0,8	0,50	Neutro
W	-0,9	0,55	Neutro
Y	-1,3	0,60	Neutro
P	-1,6	0,65	Neutro
H	-3,2	0,70	Hidrofílico
Q	-3,5	0,75	Hidrofílico
N	-3,5	0,80	Hidrofílico
E	-3,5	0,85	Hidrofílico
D	-3,5	0,90	Hidrofílico
K	-3,9	0,95	Hidrofílico
R	-4,0	1,00	Hidrofílico

5.5 Inserção de Zeros

Após a etapa de codificação da estrutura primária das proteínas, o problema de vetores de diferentes tamanhos voltou a ocorrer. Isto acontece porque a estrutura primária da proteína considera a ordem dos aminoácidos ao longo da cadeia. Ou seja, é dependente do tamanho da proteína, que como vimos na Figura 3, é bastante variável. Porém, devido à grande quantidade de informações contida na estrutura primária, a utilização da TDC se tornaria inviável, pois apesar de serem selecionados os valores mais significativos, a ordem dos aminoácidos na cadeia seria perdida.

Assim, para manter as sequências com o mesmo tamanho, foi feita a inserção de zeros, semelhante à realizada por [13], em todos os vetores com a estrutura primária, de forma a preencher as posições que faltavam até chegar ao tamanho do maior vetor da base de dados, que foi de 1014 aminoácidos. Ou seja, com a inserção de zeros, todas as proteínas passaram a possuir o mesmo tamanho, uma condição necessária para a utilização do classificador SVM.

5.6 Frequência dos Aminoácidos

Após termos selecionado as principais características e realizado a codificação da estrutura primária, inseriu-se a porcentagem de cada um dos aminoácidos da sequência proteica. Com isto, criou-se um conjunto de 20 características, cada uma delas com a frequência de um determinado aminoácido na proteína. Isto também foi feito em alguns trabalhos da literatura, como [17] e [12].

5.7 Balanceamento dos Dados

O desbalanceamento de classes é caracterizado pelo desequilíbrio na quantidade de instâncias das classes que compõem a base de dados. Neste caso, a classe majoritária possui o maior número de instâncias, e as classes minoritárias, um menor número [22] [23] [24].

Neighborhood Cleaning (NCL) é um método de *undersampling*, proposto por [25], para minimizar o impacto do problema de desbalanceamento de classes em tarefas de classificação. O método consiste em eliminar instâncias da classe majoritária que, através do algoritmo *K Nearest Neighbors* (KNN), são classificadas erroneamente pelos seus vizinhos mais próximos e que classificam instâncias da classe minoritária, aumentado a limpeza dos dados.

Devido ao desbalanceamento entre as classes da base de proteínas utilizada neste trabalho, o método NCL com 3NN (três vizinhos mais próximos) foi aplicado para o balanceamento artificial da base com o objetivo de melhorar os resultados obtidos na classificação das proteínas. O método foi aplicado duas vezes: o 1º balanceamento foi realizado com a classe majoritária das Hidrolases, e o 2º balanceamento com a classe das Tranferases, que após o 1º balanceamento ficou sendo a classe majoritária. A Tabela 5 apresenta a quantidade de instâncias originais de cada classe, após o 1º balanceamento, e por fim a quantidade que foi utilizada para os testes, após o 2º balanceamento.

Tabela 5. Distribuição das enzimas nas classes utilizadas.

Classe de Enzimas	Quantidade Original	1º Balanceamento	2º Balanceamento
Oxidoredutase	76	76	76
Transferases	120	120	70
Hidrolases	161	87	87
Liases	60	60	60
Isomerases	57	57	57
Ligases	18	18	18
Total	492	418	368

5.8 Seleção dos Melhores Parâmetros da SVM

O *kernel* mais utilizado para classificação de proteínas através de SVMs é o *Radial Basis Function* (RBF) [26]. Com a RBF, diferentemente do *kernel* linear, é possível resolver problemas, originalmente não linearmente separáveis, através do mapeamento para um espaço de maior dimensão. Nesse tipo de *kernel* o parâmetro γ (gamma), que determina a largura da curva gaussiana, precisa ser ajustado. Além deste, o parâmetro c (custo), que funciona como um tipo de tolerância aos erros presentes em um problema de classificação, também precisa ser adequado à base de dados.

No trabalho de [13] foi utilizado um algoritmo genético que selecionava quais os melhores valores para esses parâmetros. Neste trabalho foi utilizado um *script* chamado de *Grid Search* disponibilizado pelos autores do LibSvm [27] (versão de SVM utilizada nesse trabalho e que funciona integrada à ferramenta Weka). O *Grid Search*, faz uma busca exaustiva pelos melhores parâmetros, realizando a classificação diversas vezes, e verificando qual será o melhor resultado obtido. Para este trabalho, os melhores valores encontrados para os parâmetros γ (gamma) e c foram, respectivamente, 0.0034 e 3.0314.

5.9 Ferramenta de Classificação e Método de Amostragem

Para classificar os dados, foi utilizada a ferramenta WEKA, um pacote desenvolvido pela Universidade de Waikato em 1993, com o intuito de agregar algoritmos de aprendizado de máquina para mineração de dados na área de Inteligência Artificial, como por exemplo: SVM, Modelo Bayesiano, Redes Neurais, Regressão Linear, Árvores de Decisão, IB1, *Bagging*, *LogistBoot*, etc [16].

Além de disponibilizar os algoritmos de classificação, essa ferramenta também disponibiliza alguns métodos de amostragem de dados. Um método de amostragem deve ser capaz de determinar se uma hipótese vai ser suficiente para prever os dados futuros de forma eficaz. A técnica utilizada nesse trabalho foi o *k-Cross-Validation* [**?**].

5.10 Métricas de Avaliação

Para avaliar o desempenho do método proposto, foram utilizadas 4 métricas: precisão, sensibilidade, acurácia e especificidade [13].

Precisão é a taxa de instâncias corretamente classificadas como pertencentes à classe em questão, dentre todas as que foram classificadas na classe em questão, conforme a Equação 2.

$$\textbf{Precisão} = \left(\frac{VP}{VP + FP}\right) * 100 \tag{2}$$

Sensibilidade é a taxa de instâncias corretamente classificadas como pertencentes à classe em questão, dentre todas as que realmente são da classe em questão, como apresentado na Equação 3.

$$\textbf{Sensibilidade} = \left(\frac{VP}{VP + FN}\right) * 100 \tag{3}$$

Acurácia é a taxa total de instâncias corretamente classificadas e é definida pela Equação 4.

$$\textbf{Acurácia} = \left(\frac{VP + VN}{VP + VN + FP + FN}\right) * 100 \tag{4}$$

Especificidade avalia se as instâncias que não são de uma determinada classe foram realmente classificadas como não pertencentes àquela classe (Equação 5).

$$\textbf{Especificidade} = \left(\frac{VN}{VN + FP}\right) * 100 \tag{5}$$

Sendo, *Verdadeiro Positivo (VP)* o número de proteínas corretamente classificadas na classe em questão; *Falso Negativo (FN)* a quantidade de proteínas da classe analisada, erroneamente classificada; *Falso Positivo (FP)* número de proteínas que não são da classe considerada, mas que foram classificadas nesta classe; *Verdadeiro Negativo (VN)* número proteínas não pertencentes à classe considerada e que não foram classificadas nesta classe.

6 Resultados e Discussões

Os primeiros testes foram realizados com o objetivo de determinar qual o melhor valor da TDC a ser utilizado e qual o impacto da normalização dos dados nos resultados. Para isso, foram realizados testes com os dados normalizados, variando-se o número de coeficientes entre 20 e 800, com intervalo de 10, e medindo-se a precisão encontrada, como pode ser visto na Figura 5.

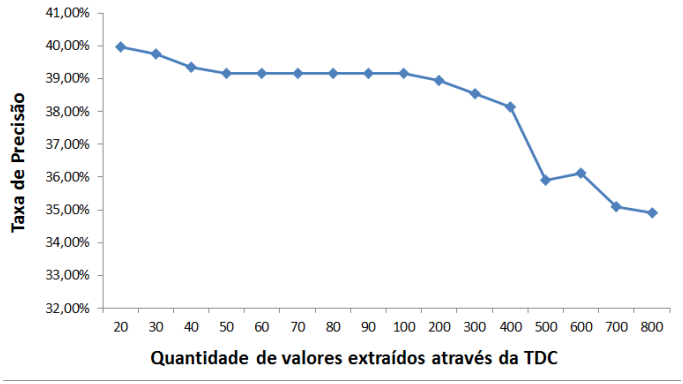


Figura 5. Resultados da utilização da TDC variando-se o número de coeficientes.

A escolha do valor máximo, 800, foi devido ao maior valor possível para utilização da TDC neste trabalho, correspondente ao número de características utilizadas, no caso 10, multiplicado pelo tamanho da menor cadeia de aminoácidos da base de dados, 80. A Figura mostra que, assim como no trabalho de [12], o melhor valor de coeficientes a ser utilizado foi 20, com precisão de 40%, e à medida que aumentamos a quantidade de coeficientes utilizados, verificou-se que a taxa de precisão caiu.

Uma vez escolhido o valor do coeficiente da DTC, foram realizados testes para verificar o quanto a normalização dos dados afeta os resultados, extraindo os 20 valores através da TDC e calculando sua precisão. Foram criados dois arquivos de entrada com as características descritas na Seção 5.1, um com os dados sem normalização e outro com os dados normalizados. Verificou-se que a precisão encontrada com os dados não normalizados foi de 38,34%, ao passo que com os dados normalizados foi de 39,96%. Assim, optou-se por utilizar os dados normalizados na metodologia final, visto que houve um ganho de 1,62%.

Após ter escolhido o valor da TDC, o nosso objetivo foi verificar qual a melhor combinação de metodologias para a predição de função de proteínas. Para cada teste, foi utilizado o *Grid Search* para a seleção dos melhores parâmetros da SVM. Optamos por não aplicar o passo do balanceamento dos dados, descrito na Seção 5.7 nesses testes, para que seja possível discutir posteriormente o impacto dessa técnica nos resultados finais. De forma a simplificar a visualização dos resultados, apresentaremos apenas o gráfico comparativo para a métrica de precisão (Fig. 6).

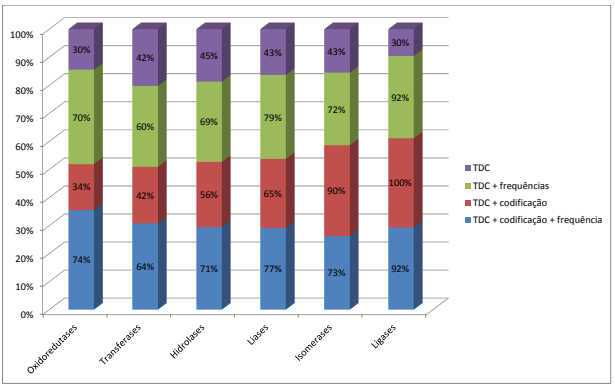


Figura 6. Resultados da precisão utilizando-se a metodologia proposta.

Primeiramente avaliou-se a precisão considerando apenas a TDC sobre as 10 características extraídas do banco Sting_DB. Nota-se pela Figura que, com apenas as 10 características selecionadas, o classificador obteve um desempenho máximo de 45% para as hidrolases.

Testando o classificador com a TDC sobre 10 características, acrescido da frequência dos aminoácidos, o classificador melhorou de forma significativa a predição de todas as classes, com um aumento máximo de 200% para as ligases e mínimo de 42% para as transferases. Em média o ganho da precisão entre estes dois testes (TDC e TDC + frequência dos aminoácidos) foi de 90%.

Testamos também o impacto que a codificação (apresentada na Tabela 4) traria ao classificador. Pela Figura 6, percebe-se que houve um ganho máximo de 233% para as ligases e que a precisão das transferases manteve-se constante. O ganho médio entre estas duas metodologias (TDC e TDC + codificação) foi de 67%.

O teste final foi realizado para verificar o quanto a TDC + codificação + frequência poderiam melhorar o classificador, visto que os dois testes anteriores (TDC + frequência e TDC + codificação) melhoraram a predição das classes de forma diferente. Observa-se que o ganho médio no valor da precisão foi de 92%, tendo os maiores ganhos nas classes das oxidoredutases, com 149%, e nas ligases com 200%. Os menores ganhos foram para as transferases, com 52%, e hidrolases com 58%.

A partir desses testes, analisou-se as quatro métricas avaliadas neste trabalho. A Tabela 6 mostra que, mesmo sem aplicarmos o balanceamento dos dados, ou seja, usando todas as 492 proteínas da base, já obtivemos resultados significativos, com média a partir de 71%

para 3 das 4 métricas avaliadas (exceto para a sensibilidade que teve média de 67%). Isto já representa uma melhora dos resultados em relação aos trabalhos relacionados, que possui uma precisão máxima de 70,06% (Fig. 7). Quando há uma desbalanceamento entre as classes, o classificador tende a classificar na classe majoritária as instâncias pertencentes à classe minoritária. Ou seja, a sensibilidade das classes minoritárias tende a ser baixa, o que pode ser observado na Tabela 6 para as classes liases, isomerases e ligases. Além disso, a quantidade de falsos positivos (FP) para as classes majoritárias tende a ser alta, o que reduz a taxa da especificidade.

Esse problema acontecia principalmente com as classes de transferases e hidrolases que possuiam um número muito maior de instâncias do que as outras classes (Tabela 5). Por isso, foram feitos dois balanceamentos, respectivamente para essas classes majoritárias.

Tabela 6. Medidas de desempenho **sem** balanceamento dos dados, considerando-se TDC+codificação+frequência.

	Precisão	Sensibilidade	Acurácia	Especificidade
Oxidoredutases	74%	66%	71%	76%
Transferases	64%	73%	66%	58%
Hidrolase	71%	79%	73%	68%
Liase	77%	57%	70%	83%
Isomerase	73%	65%	70%	75%
Ligase	92%	61%	78%	94%
Média	75%	67%	71%	76%

Feito isso, foram realizados testes utilizando toda a metodologia proposta, ou seja, com TDC+codificação+frequência+balanceamento. A Tabela 7 apresenta estes resultados.

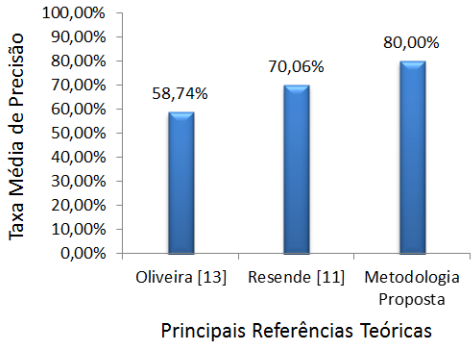
Tabela 7. Medidas de desempenho **com** balanceamento dos dados, considerando-se TDC+codificação+frequência.

	Precisão		Sensibilidade		Acurácia		Especificidade	
	1º bal.	2º bal.	1º bal.	2º bal.	1º bal.	2º bal.	1º bal.	2º bal.
Oxidoredutase	65%	52%	88%	70%	66%	54%	62%	20%
Transferases	66%	80%	70%	85%	71%	76%	57%	83%
Hidrolases	90%	93%	82%	80%	86%	88%	91%	94%
Liases	64%	77%	57%	53%	62%	70%	70%	83%
Isomerases	88%	79%	67%	65%	78%	75%	91%	82%
Ligases	92%	100%	61%	61%	78%	81%	94%	100%
Média	77%	80%	69%	71%	73%	74%	77%	77%

Se compararmos as Tabelas 6 e 7, nota-se que todas as métricas utilizadas, considerando-se os valores médios, obtiveram uma melhora, tanto após o primeiro balanceamento para a classe das Hidrolases, quanto para o segundo na classe das Transferases.

Após obtermos os resultados para cada métrica, foi feita uma comparação dos resultados com alguns dos principais trabalhos da literatura que utilizaram essa mesma base de dados. A Figura 7 apresenta a média da precisão para os trabalhos de [13] e [12]. Pode-se observar que a média da metodologia proposta neste trabalho supera a média de precisão dos trabalhos anteriores. É importante ressaltar que, apesar de o trabalho descrito em [18] estar relacionado com predição de função de proteínas, as classes analisadas pelos autores não foram as mesmas estudadas neste projeto, e, portanto, não há como compará-los.

Figura 7. Comparação da métrica precisão com outros trabalhos.



7 Considerações Finais

Neste trabalho foi apresentada uma metodologia para predição de função de proteínas, que combina recursos da matemática (TDC), aprendizado de máquina (SVM), dados da estrutura primária e secundária e diversas características de proteínas disponibilizadas pelo banco de dados Sting_DB. Os resultados de cada etapa da metodologia foram comparados entre si, mostrando os ganhos obtidos em cada teste.

Os resultados confirmam a importância da normalização dos dados, a relevância de se adicionar características ao classificador (o que ofereceu os melhores ganhos) e a necessidade do balanceamento das classes.

Como trabalhos futuros, pretende-se implementar uma técnica de seleção de atributos, tal como Algoritmos Genéticos, para a seleção automática das características do banco de dados Sting_DB. Além disto, pretendemos adicionar mais características extraídas de outros banco de dados, tais como o Pfam e Interpro.

Agradecimentos

Os autores agradecem o apoio financeiro recebido da Fundação de Amparo à Pesquisa do Estado de Minas Gerais-FAPEMIG (Projeto APQ-01565-12) e ao Centro Nacional de Supercomputação (CESUP) da Universidade Federal do Rio Grande do Sul (UFRGS).

Bibliografia

[1] F. Prosdocimi, G. Coutinho, E. Ninnecw, A. F. Silva, A. N. dos Reis, A. C. Martins, A. C. F. dos Santos, A. N. Júnior, and F. Camargo Filho, "Bioinformática: manual do usuário," *Biotecnologia Ciência & Desenvolvimento*, vol. 29, pp. 12–25, (2002).

[2] V. G. Bittencourt, "Aplicações de técnicas de aprendizado de máquina no reconhecimento de classes estruturais de proteínas," Master's thesis, Universidade Federal do Rio Grande do Norte, (2005).

[3] N. Nadzirin and M. Firdaus-Raih, "Proteins of unknown function in the protein data bank (pdb): An inventory of true uncharacterized proteins and computational tools for their analysis," *International Journal of Molecular Sciences*, vol. 13, no. 10, pp. 12 761–12 772, (2012). [Online]. Available: http://www.mdpi.com/1422-0067/13/10/12761

[4] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne, "The protein data bank," *Nucleic acids research*, vol. 28, no. 1, pp. 235–242, Jan. (2000).

[5] B. Alberts, D. Bray, A. Johnson, J. Lewis, M. Raff, K. Roberts, and P. Walter, *Fundamentos da biologia celular: uma introdução à biologia molecular da*

célula. Artmed, (2002). [Online]. Available: http://books.google.com.br/books?id=25X4QwAACAAJ

[6] P. D. Dobson and A. J. Doig, "Predicting enzyme class from protein structure without alignments," *Journal of Molecular Biology*, vol. 345, no. 1, pp. 187 – 199, (2005).

[7] S. de Apoio ao Ensino, "Bioquímica: Proteínas," Mar. (2006), acesso em: 18 Abr. 2013. [Online]. Available: {http://www.iesde.com.br/pai/arquivos/EM_1S_BIO_003.pdf}

[8] A. Lehninger, D. L. Nelson, and M. M. Cox, *Lehninger Principles of Biochemistry*. Macmillan, (2008).

[9] V. Vapnik, *The Nature of Statistical Learning Theory (Information Science and Statistics)*, 2nd ed. Springer, Nov. (1999). [Online]. Available: http://www.amazon.com/exec/obidos/redirect?tag=citeulike07-20\&path=ASIN/0387987800

[10] A. C. Lorena, André, and C. P. L. F. de Carvalho, "Uma introdução às support vector machines," *Revista de Informática Teórica Aplicada*, vol. 14, no. 2, pp. 43–67, (2007).

[11] C. wei Hsu, C. chung Chang, and C. jen Lin, "A practical guide to support vector classification," National Taiwan University, Taiwan, Tech. Rep., (2010).

[12] S. R. M. Oliveira, L. C. Borro, M. E. B. Yamagishi, A. L. Mancini, J. G. Jardine, I. Mazoni, E. H. d. Santos, R. H. Higa, P. R. Kuser, and G. Neshich, "Predicting enzyme class from protein structure using bayesian classification." *Genet Mol Res*, vol. 5, no. 1, pp. 193–202, (2006).

[13] W. Resende, R. Nascimento, C. Xavier, I. Lopes, and C. Nobre, "The use of support vector machine and genetic algorithms to predict protein function," in *Systems, Man, and Cybernetics (SMC), 2012 IEEE International Conference on*, oct. (2012), pp. 1773 –1778.

[14] G. Neshich, W. Rocchia, A. L. Mancini, M. E. B. Yamagishi, P. R. Kuser, R. Fileto, C. Baudet, I. P. Pinto, A. J. Montagner, J. F. Palandrani, J. N. Krauchenco, R. C. Torres, S. Souza, R. C. Togawa, and R. H. Higa, "Java protein dossier: a novel web-based data visualization tool for comprehensive analysis of protein structure," *Nucleic Acids Research*, vol. 32, pp. 595–601, (2004).

[15] A. G. Murzin, S. E. Brenner, T. Hubbard, and C. Chothia, "Scop: a structural classification of proteins database for the investigation of sequences and structures." *Journal of molecular biology*, vol. 247, no. 4, pp. 536–540, Apr. (1995).

[16] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The weka data mining software: an update," *SIGKDD Explor. Newsl.*, vol. 11, no. 1, pp. 10–18, Nov. (2009). [Online]. Available: http://doi.acm.org/10.1145/1656274.1656278

[17] A. L. D. Rossi and M. A. de O. C. Brunetto, "Métodos de codificação de proteínas para uso com redes neurais artificiais," *X Congresso Brasileiro de Informática em Saúde*, (2006).

[18] U. M. Dias, "Predição da função das proteínas sem alinhamentos usando máquinas de vetor de suporte," Master's thesis, Universidade Federal de Alagoas, (2007).

[19] N. Ahmed, T. Natarajan, and K. Rao, "Discrete cosine transfom," *IEEE Transactions on Computers*, vol. 23, no. 1, pp. 90–93, (1974).

[20] W. R. Weinert and H. S. Lopes, "Aplicação de um sistema neural ao problema de classificação de proteínas," *VI Congresso Brasileiro de Redes Neurais*, pp. 85–90, (2003).

[21] J. Kyte and R. F. Doolittle, "A simple method for displaying the hydropathic character of a protein." *Journal of molecular biology*, vol. 157, no. 1, pp. 105–132, May 1982. [Online]. Available: http://view.ncbi.nlm.nih.gov/pubmed/7108955

[22] P.-N. TAN, M. STEINBACH, and V. KUMAR, "Introdução ao data mining: Mineração de dados." *Editora Ciência Moderna LTDA*, p. 350, (2009).

[23] P. C. de Oliveira Brandão, "A análise da influência do balanceamento de classes em máquina de vetores de suporte," Master's thesis, Pontifícia Universidade Católica de Minas Gerais, (2013).

[24] G. E. de Almeida Prado Alves Batista, "Pré-processamento de dados em aprendizado de máquina supervisionado," Master's thesis, Universidade de São Paulo, (2003).

[25] J. Laurikkala, "Improving identification of difficult small classes by balancing class distribution," Master's thesis, University of Tampere, (2001).

[26] B. Martin, "Radial basis functions: theory and implementations," (2003).

[27] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, pp. 27:1–27:27, (2011), software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.